bioRxiv preprint doi: https://doi.org/10.1101/030718; this version posted November 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

On characterizing membrane protein clusters with *model-free* spatial correlation approaches

A. Shivanandan, J. Unnikrishnan, A. Radenovic

Abstract

Spatial aggregation or clustering of membrane proteins could be important for their functionality, e.g., in signaling, and nanoscale imaging can be used to study its origins, structure and function. Such studies require accurate characterization of clusters, both for absolute quantification and hypothesis testing. A set of *model-free* quantification approaches — *free* of specific cluster models— have been proposed for this purpose. They include the radius of maximal aggregation r_a obtained from the maxima of the empirical Besag $L(r) - r$ function as an estimator of cluster size, and the estimation of various cluster parameters based on an exponential approximation for the Pair Correlation Function(PCF). However, the parameter identifiability and bias and scaling due to their *model-free* nature are not clear. In practice, the clusters might exhibit specific patterns, and the behavior of these estimators in such cases must be studied. Here, we theoretically analyze these approaches for a set of cluster models, and obtain information about their identifiability and bias. We find that the *ratio* between r_a and true cluster size depends on both the true size as well as the number of clusters per unit area, or other corresponding parameters, in a model-dependent manner. In particular, r_a scales with respect to the true size by a factor that can be arbitrarily large, depending on models and parameter values. For the method based on PCF approximation, for most models we analyzed, the ratios between approximate and true model parameters were found to be constants that depend only on models and independent of other parameters. For the models analyzed, this ratio was within $\pm 100\%$. Our theoretical approach was validated by means of simulations. We also discuss some general issues in inference using second-order spatial properties. While precision could also be key, such information on identifiability and accuracy provides clarity on estimation, can lead to better inference, and can also fuel more accurate method development.

Introduction

In cell biology and elsewhere, spatial aggregation or clustering is an interesting phenomenon, possibly with a functional role — e.g., the behavior of membrane proteins to form sub-micrometer clusters could be important for their functionality, such as in signaling (1–4). The origins, structure and function of spatial heterogeneity in membrane proteins are only being studied. Spatial location information, available from fluorescence and electron microscopic imaging, and recently from sub-diffraction limited fluorescence imaging such as Single Molecule Localization Microscopy(SMLM)

bioRxiv preprint doi: https://doi.org/10.1101/030718; this version posted November 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

techniques (5–7), are key to such studies (8–11). Accurate characterization of clustering — the strength, scale and density of clustering is an important part of these studies, whether for relative comparison between different systems, perturbation conditions and to test hypothesis (such as the relative importance of lipid rafts and actin cytoskeleton in membrane protein clustering (12), or the possible mechanisms of early T-cell signaling (8)), or even for absolute quantification (such as the size of clusters in a particular cell type in a particular condition, and the number of molecules in them).

A number of methods have been used to characterize the clusters from imaging data (13–15). While most of these were aimed at characterizing membrane protein clusters, many of them were used to characterize other systems in the cell that exhibit clustering. The methods can be broadly categorized into two: (1) clustering or segmentation to identify the clusters, followed by their characterization; and (2) spatial statistics approaches based on a second-order spatial summary statistic such as Besag $L(r) - r$ function or the Pair Correlation Function $g(r)$. These second-order functions can be used for comparison of clustering at different scales and between different experimental systems and perturbations, and estimators based on these functions can be used for ensemble cluster parameter estimation. In general, they have a few advantages over many of the segmentation approaches: they can detect interactions at multiple spatial scales, can work with both dense and sparse point patterns, often have direct physical interpretations (16), and are amenable to rigorous extensions incorporating error models, crucial in the case of nanoscale imaging (16–18). Also, in the case of SMLM, the notion of spatial point patterns align well with the nature of its point localization readout. In practice, a major convenience of using such methods have been that they estimate ensemble functions at different scales and the various cluster parameters for a whole dataset, making comparative studies easy in systems where variability within cluster sizes are not important.

The simplicity in parameter estimation is in no small part aided by the *model-free* nature of some of these approaches. Two spatial statistics based estimators of cluster parameters based on these functions widely reported in the nanoimaging and protein cluster analysis literature, 1) the radius of maximal aggregation r_a (13, 15, 18–27), the radius value corresponding to the maxima of the empirical $L(r) - r$ function, as an estimator of cluster size(length scale); and 2) the *model-free* functional approximation of $g(r)$ as an exponential function (9, 16, 28–32), leading to estimators of cluster size, amplitude or strength and number of molecules per cluster, are not concerned with the underlying spatial distributions, such as the shapes of clusters and the distribution of molecules in them. Effects due to differences in underlying spatial distribution are either ignored or approximated, effectively making the estimation process free of underlying cluster processes, or *model-free*. The *model-free* nature of these functions vary — r_a does not contain any model of clustering whereas the PCF approximation is a generic function independent of specific models.

However, clusters observed through bio-imaging could be of different shapes, depending on the underlying physical mechanism. In the case of SMLM imaging, e.g., the clusters formed due to photoblinking are reported to have a Gaussian (9) or Cauchy peak shape (32), depending on the photon count distribution within the cluster. It is plausible to model internalization in circular or spherical bodies with a hard-core process (a disk in 2D). Analysis methods often assume Gaussian shapes for membrane clusters (14, 18). (33, 34) have suggested modeling membrane protein distributions using 2D-Ising model, to account for phase transitions and criticality. It is not clear how the parameter estimation approaches that are *model-free* are biased or scaled due to these different underlying true cluster processes. Also, such *model-free* approximations also raise the

bioRxiv preprint doi: https://doi.org/10.1101/030718; this version posted November 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

question of identifiability: e.g., can the size (i.e. length scale) parameter of model-free approaches be mapped exclusively to the size parameter of the true process, independent of other parameters, such as number of clusters per unit area or cluster density or amplitude? If the estimated size parameter is dependent on both the size and amplitude parameters of the underlying true process, one must account for it during the comparative analysis of cluster sizes, as it may not accurately reflect the true differences in size, being affected by amplitudes as well. Other point pattern based parametric methods (24, 35) also have to deal with similar issues. The influence of shape and geometry in estimation is observed in other fluorescence based technologies as well (36).

Some clues have been obtained from simulation studies. Kiskowski et al (19) studied the relation between the true radius of disk clusters R and estimates of r_a by means of simulations, and derived important insights — such as $R \leq r_a \leq 2R$, and a qualitative dependency of r_a on separation between clusters. However, since the study was based on simulations, with a limited set of parameters and models (only disk clusters), the understanding is limited, and the possibilities of generalization are not clear. Lagache et al (24) performed a theoretical analysis of a similar estimator — maxima of the K-function normalized with its variance — for disk shaped clusters, and reported a simpler, constant relation $r_a/R = 1.3$. Such a relationship would have been convenient, however its generality in terms of models and parameters is not clear. No studies of the bias introduced by the approximate model of $g(r)$ has yet been reported, to the best of our knowledge.

Note that the accuracy or bias of an estimator cannot be improved by repeated measurements, unlike its precision. By definition, bias affects absolute quantification, e.g. an estimator of single molecule counts that is biased by 50% lower than the true counts affects absolute quantification. The same is the case regarding their use as relative comparisons: the need to account for biases might become important if (1) the parameters are not separately identifiable or (2) involves scaling that are model dependent and the comparisons involve different models.

In this work, we explore, with theoretical rigor, the bias in parameter estimation and the questions of identifiability introduced by these *model-free* approaches. We consider a number of spatial cluster processes whose theoretical $g(r)$ and $L(r) - r$ are known, and then derive the relation between the parameters of the *model-free* approaches (such as r_a) and the true process parameters (e.g., the cluster size parameter r_t). We find that, in general, for a large class of clustered point patterns, the ratio p of the radius of maximal aggregation r_a and the size parameter of the true process $r_t(p = r_a/r_t)$ can be derived as an implicit function of two cluster parameters: r_t and the number of clusters per unit area (κ). We also find that it possible to derive a theoretical lower bound for p, given a cluster model following some basic assumptions. We validate the theoretical results with simulations. We also perform similar analysis for the statistic presented in (24), to report a more complex relationship between the true cluster size and the estimator. Then, we investigate the bias due to the exponential approximation model of $g(r)$, for all the models listed. By minimizing the Least Square Distance between the true and approximate PCFs, we obtain scaling laws between the approximate model and the true model parameters, and validate the approach by simulations. The extension for other cluster models are straightforward.

bioRxiv preprint doi: https://doi.org/10.1101/030718; this version posted November 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Materials and Methods

Background definitions

For a spatial point pattern in 2D-space, Ripley's K-function is defined (37–39) as

$$K(r) = \frac{1}{\rho}\mathsf{E}[M(r)] \tag{1}$$

where ρ is the spatial density (average number of points per unit area), and $M(r)$ is the number of other events within distance r of a randomly chosen event. The Besag $L(r) - r$, a measure of cluster strength at r, is then given by

$$L(r) - r = \sqrt{\frac{K(r)}{\pi}} - r, \tag{2}$$

and the Pair Correlation Function by

$$g(r) = \frac{K'(r)}{2\pi r}. \tag{3}$$

Alternative but equivalent definitions of PCF starting with the notion of spatial autocorrelation are also possible (9).

The radius of maximal aggregation, $r_a = \arg\max_r L(r) - r$.

The function $g_a(r) = 1 + a\exp(-r/d)$ has been proposed as a functional approximation for the Pair Correlation Function (PCF) of "2D-system of clusters with no predefined shape" (9, 28). The parameter a is the amplitude, a measure of point density in the clusters, and d, the correlation length, gives the radius of the cluster (9). For the PCF $g(r)$, the average number of points per cluster can then be obtained as

$$N^{cluster} = 1 + \rho\int_0^\infty (g(r) - 1)2\pi r dr, \tag{4}$$

which is equal to $N_a = 2\pi a d^2\rho$ in the case of $g_a(r)$, where ρ is the average density of points in the area of analysis(9).

Theoretical expressions for $g(r)$ and $L(r) - r$

In order to derive the theoretical expressions for $L(r) - r$ and $g(r)$ for different cluster models, it is useful to focus on a class of spatial cluster processes, known as Poisson cluster processes, or Neyman-Scott processes(details in (38, 39)), which are generated in the following way. First, a set of *parent* points are created, following a spatial Poisson process (complete spatial randomness) with density (intensity) κ. Then, S number of points are distributed around each *parent* point according to the i.i.d bivariate PDF $f_{pdf}(.)$, S following some i.i.d distribution with mean μ. These *offspring* points form the clustered point pattern. Such simple spatial cluster models that consider different shapes of clusters provide a starting point for the theoretical analysis of estimators. The analysis of Ising model in the later sections provide a more physical example.

Assuming $f_{pdf}(.)$ to be radially symmetric, let the PDF of the distance r between two offspring points within a cluster is given by $h_d(r)$ and its Cumulative Distribution Function (CDF) by $H_d(r)$. Then(38):

bioRxiv preprint doi: https://doi.org/10.1101/030718; this version posted November 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$K_{clust}(r) = \pi r^2 + \frac{\mathsf{E}[S(S-1)]}{\kappa\mu^2} H_d(r). \tag{5}$$

The density of the point pattern will be $\mu\kappa$. When $S \sim Poiss(\mu)$, since $E[S(S-1)] = \mu^2$ for Poisson distribution, (5) reduces to

$$K_{clust,Poisson}(r) = \pi r^2 + \frac{1}{\kappa} H_d(r). \tag{6}$$

The derivation in case of other distributions for points per cluster is straightforward. In the case of geometric or exponential distribution of S, behavior often observed in nanoimaging (40–42), $\frac{\mathsf{E}[S(S-1)]}{\mu^2} \to 2$ for $\mu \gg 1$.

Note that H_d, being the CDF, is monotonic and non-decreasing. The corresponding PCF $g(r) = \frac{K'(r)}{2\pi r}$ becomes:

$$g_{clust,Poisson}(r) = 1 + \frac{1}{2\pi\kappa}\frac{h_d(r)}{r} \tag{7}$$

The PCF and K-function for different cluster shapes are given in Table 1, and the shapes of their PDFs are given in Supporting Material. Note that disk clusters contain points distributed uniformly at random within a circle (disk), a process known as Matérn cluster process in spatial statistics (the case of Gaussian cluster shapes is known as Thomas process). Also note that r_t is defined differently for different cluster models: for a disk cluster, $r_t = R$, the true cluster radius, whereas for Gaussian clusters, we set $r_t = \sigma$, the true standard deviation (the full list can be found in Table 1). We also add the physical Ising model to the compilation, since it is one of the models that has been proposed for membrane protein clustering (16), even though it is not a Neyman-Scott process. Also, note that the exponential approximation $g_a(r)$ has the same shape as the variance Gamma function model(varGamma) in Table 1, pointing at the non-uniqueness of $g(r)$ shapes and the difficulty of identifying cluster models from data based on their PCF shapes.

Effect of background To model a monomer fraction or background, a spatial Poisson distributed monomer point pattern can be superimposed to a purely clustered process, such that the purely clustered fraction of points is β. The resulting K-function and PCF can be obtained using the expression for superposition of two independent point processes (38). In the case of a clustered process with $g(r) = 1 + Bv(r)$, superposition with such a background process results in the PCF:

$$g(r) = 1 + B_e v(r), \tag{8}$$

where $B_e = B\beta^2$, β being the purely clustered fraction (38). Expressions for $K(r)$ and $L(r) - r$ undergo similar scaling in parameter. It can be noted that the shape of the function remains the same as the purely clustered process, the change in parameter B being the only change, again pointing at the non-uniqueness of PCF shapes, and the quadratic effect of background on the function(note the effect on (4)).

Simulation and analysis details

All simulations were done in `R`, using the `spatstat` library (43). Simulations of cluster processes were done with standard library functions, such as `rThomas` and `rMatClust`. Parameter estimation

bioRxiv preprint doi: https://doi.org/10.1101/030718; this version posted November 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

by minimum contrast method was done using `kppm` function, and using parameters "Thomas" and "VarGamma". Analytical derivations were performed with the help of symbolic algebra software [Mathematica(Wolfram Research, USA)].

Results

Estimation based on radius of maximal aggregation

Theoretical expressions for radius of maximal aggregation Here we analyze the relationship between the radius of maximal aggregation, defined as $r_a = \arg\max_r L(r) - r$, as a function of true cluster parameters, for the class of clustered point patterns with K-functions of the form

$$K(r) = \pi r^2 + \frac{1}{A} H(r) \tag{9}$$

where $h(r) = H'(r)$ and $A > 0$, such as the ones introduced in the Methods section and Table 1. Then, $L'(r_a) - 1 = 0$, $L'(r_a) = 1 \implies K'(r_a)^2 = 4\pi K(r_a)$, using (2). Substituting in (9), we obtain

$$A = \frac{h(r_a)^2}{4\pi(H(r_a) - r_a h(r_a))}. \tag{10}$$

That is, r_a depends on A in general, as A is not a parameter of H and h. (10) can be used to obtain a relation between $p = \frac{r_a}{r_t}$ for all the models listed in Table 1, where r_t is the cluster size parameter of the true process. The results are given in Table 2(more details in Supporting Material). It is possible to write the relationship $f(p) = r_t^2 \kappa$ for all the Poisson cluster processes discussed. In the case of the Ising process, the corresponding relationship is of the form $f(p) = \frac{1}{2\pi} a_I^{-1} \xi^{1/4}$. The derivation of p in the case of a power-law PCF is given in Supporting Material.

Note that the expression for p (and hence r_a) is independent of the number of points per cluster (μ) if the expressions for K-functions are independent of it. Figure 1 shows the contour plot of p *vs* κr_t. Thus, we can establish that theoretically, the ratio between the radius of maximal aggregation and the true size parameter is dependent on both the true size parameter as well as the number of clusters per unit area. In fact, the singularity at $H(r_a) - r_a h(r_a) = 0$ provides a minimum bound for p for all the models analyzed, and is also shown in Table 2(see Appendix for a proof). The lower bound so obtained is a fundamental characteristic of the cluster model's theoretical $L(r) - r$ functions. The existence of a lower bound for r_a for any cluster model with K-function of the form in (9) can be proved theoretically given some basic assumptions on $h(r)$(see Appendix).

The $p = r_a/r_t$ for different processes cannot be directly compared, as the size parameter of the true process r_t is defined differently for them. Now, it can be seen that the lower bound for p is model dependent. For the disk model, e.g., $1.29564 < p < 2$, whereas, for Gaussian model, $2.24181 < p < \infty$. A more comparable measure would be r_q, the (true) scale at which q fraction of the points are expected to lie for a particular distribution, typically obtainable in the form $r_q = u_q r_t$, such as the case of $r_{.95} = 2\sigma$ in the case of 1D Gaussian distribution. This would correspond to the ratio $p_q = r_a/r_q = p/u_q$, and $f(p_q u_q) u_q^2 = \kappa r_q^2$. Considering the case $q = 95\%$, the values for $u_{.95}$ and the lower bounds for $p_{.95}$ for different distributions is given in Supporting Material, and the plot $p_{.95}$ *vs* $\kappa r_{.95}^2$ is shown in Figure 1c. It can be seen that $p_{.95}$ is dependent on both the model as well as both the number of clusters per unit area and the true cluster size.

The systematic relationship established between p (or $p_{.95}$),A and r_t, clarifies the bias and identifiability issues in estimation. The results agree with (19), and provides a tighter (theoretical)

bioRxiv preprint doi: https://doi.org/10.1101/030718; this version posted November 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

lower bound for disk clusters. The approach can also explain the qualitative influence of inter-cluster distance on r_a, observed by (19), through the dependency of p on κ, along with the relative influence of κ and r_t. The dependency of p on other cluster parameters and the cluster model means that the estimator could be a poor choice as a comparison tool between different experiments, if there is a possibility that the cluster model, κ or r_t are different — unless there is a large enough difference in estimated r_a for different experiments.

Validation with simulations To establish the validity of the theoretical derivation obtained in previous section (shown in Table 2) we performed a Monte Carlo simulation study. In addition to information about the accuracy of radius of maximal aggregation(the subject of the theoretical study), it also provides information about its precision as an estimator.

Clustered point patterns, belonging to either Gaussian or disk clusters, were simulated in a unit square, for varying κ and r_t. The theoretical value of p for a given κ and r_t were obtained by solving the analytical expressions in Table 2, and was compared to $\hat{p} = \hat{r}_a/r_t$. $\hat{r}_a$ was obtained from the empirical maximum of the $L(r) - r$ curves. The results are shown in Figure 2(also see Supporting Material Figure 1). The mean value of $\hat{p}$ from simulations broadly agree with the theoretical results, though the deviation increases with increasing κr_t^2(see also the Mean Squared Error in Figure 2c,d). This is probably the result of increasing number of clusters per unit area (increasing κ) or having larger clusters within the unit square used in the simulations (increasing r_t), both resulting in overlapping clusters, resulting in deviations from theoretical framework based on a particular cluster model. In fact, it can be seen that the deviation is most influenced by increasing radius(Figure 2c,d).

Case of normalized K-function In (24), the normalized statistic $\tilde{K}(r,n)$ was proposed, given by $\tilde{K}(r,n) = \frac{K(r,n)-\pi r^2}{\sqrt{\mathrm{Var}(K(r,n))}}$, where

$$\mathrm{Var}(K(r,n)) = \frac{2\pi A r^2}{n^2}\left(\frac{\pi r^2\left(\frac{0.0132nPr}{A} - 1\right)}{A} + \frac{0.305Pr}{A} + 1\right), \tag{11}$$

where A is the area and P the perimeter of the observation window, and n the number of points.

For disk process, they use, similar to the expression in Table 1:

$$K(r,n) = \frac{2\beta^2\left(\frac{r^2\cos^{-1}\left(\frac{r}{2R}\right)}{R^2} - \frac{r\left(r^2+2R^2\right)\sqrt{1-\frac{r^2}{4R^2}}}{4R^3} + \sin^{-1}\left(\frac{r}{2R}\right)\right)}{\pi\kappa} + \pi r^2, \tag{12}$$

where κ is the number of clusters per unit area, and β the clustered fraction.

The radius of maximal aggregation $\tilde{r}_a$ for $\tilde{K}(r,n)$ is then obtained by setting $\frac{\partial\tilde{K}(r,n)}{\partial r} = 0$. Using numerical approaches, they obtained the constant relation $\tilde{r}_a = 1.3R$.

In our hands, $\frac{\partial\tilde{K}(r,n)}{\partial r} = 0$ for a square observation window(for simplicity) resulted in a more nuanced situation, as shown in Figure 3(details in Supporting Material). We found that $\tilde{p} = \tilde{r}_a/r_t$ depends on the number of points n and the ratio m between the side length of the square observation window and the true size parameter r_t, and converges to a maximum value at large m, which is approximately equal to the minimum values obtained in the case of r_a based on $L(r) - r$. For example, in the case of clusters with $R = 20nm$ with an area of analysis of size $10\mu m$, then $m = 500$, and $\tilde{p}$ is close to the maximum value (Figure 3), and hence a constant(1.296 in the case of disk

bioRxiv preprint doi: https://doi.org/10.1101/030718; this version posted November 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

clusters, approximately equal to the factor of 1.3 obtained in (24)). On the other hand, if the area of analysis was smaller, say $1\mu m$, then $m = 50$, and $\tilde{p}$ depends critically on n(Figure 3). The dependency of $\tilde{p}$ on n, in contrast with p in the case of $L(r) - r$, is because $\sqrt{\mathrm{Var}(K(r,n))}$ is non-linearly dependent on n, whereas the expression for $K(r,n)$ (and $L(r) - r$) is independent of n. Note that $\tilde{p}$ is independent of κ and β, unlike the case of r_a and $L(r) - r$ presented in the previous section.

Estimation based on Pair Correlation Function

We now consider another estimator that has been suggested for estimating cluster parameters, the approach based on fitting Pair Correlation Functions. As discussed in the section Methods, the theoretical PCF is not unique for a cluster model, and its signature shape and sensitivity are often not sufficient to identify the models(Figure 4), not the least because each experiment provides a realization of a stochastic process, with the observed statistic approaching the theoretical one only as $n \to \infty$. Model selection based on Monte Carlo (MC) rank tests (38, 39) — ranking the empirical statistic value among the values of the statistic from MC simulations based on estimated parameters — based on PCF or the related K or $L(r) - r$ functions is not sound, if the same function was used for parameter estimation(38). The standard method in this case is to perform MC rank tests with a statistic that is different from the one that was used for parameter estimation, e.g., the nearest neighbor distribution function if the PCF was used for estimation. However, the approach is known to have low statistical power (39), and we too had similar experience during preliminary attempts to identify the cluster models from simulations and SMLM data (results not shown). Therefore, the *model-free* functional approximations such as $g_a(r) = 1 + a\exp(-r/d)$, proposed as part as the PC-PALM method, have much appeal.

Here, we derive a measure of bias in parameters introduced by this approximation, given a true model. We aim to find the relations $m = d/r_t$, $n = a/a_t$ and $l = N_a/N_t$, given a true model for the PCF in the form $f(r) = 1 + a_t v(r, r_t)$. Here, N_a and N_t are the average number of points per cluster corresponding to the approximate model and the true model respectively, as per (4). Given a specific model for $f(r)$, we find the relation between parameters in the case of the fit that provides the minimum (Least) Squared Error E, i.e.,

$$(\hat{a}, \hat{d}) = \arg\min_{a,d} E = \arg\min_{a,d} \int_0^{r_m} (f(r) - g_a(r))^2 dr. \tag{13}$$

Note that the Least Squares criteria was used in original PC-PALM papers for parameter estimation (9, 44). If E has a minima at $(\hat{a}, \hat{d})$, then $\frac{\partial E}{\partial a} = 0$ and $\frac{\partial E}{\partial d} = 0$ at $(\hat{a}, \hat{d})$, which can be solved to obtain expressions for $(\hat{a}, \hat{d})$. Measures of m, n and l can then be found using these.

We were able to obtain measures of m, n and l for all the cluster models described in Table 1, and the results are shown in Table 3 and the best fit PCFs can be seen in Figure 5a(details in Supporting Material). The $m_{.95}$ values: $m_{.95} = d/r_{.95} = m/u_{.95}$, $r_{.95}$ being the scale at which 95% of points are expected to lie, can also be obtained as constant scalar values, given by .63,.82,.38 and .28, for Gaussian, disk, Cauchy and varGamma models respectively.

For example, in the case of Gaussian shaped clusters, with the PCF given in Table 1, we obtain, for $r_m > 6\sigma$, $m = d/\sigma \approx 1.54, n \approx 1.26, l \approx 1.48$, with $m_{.95} = .63$. The parameters can be either *upscaled* or *downscaled* — e.g., the number of molecules per cluster is overestimated by 50% by using $g_a(r)$ for estimation, whereas in the case of Ising process, it is underestimated by 40%. The

bioRxiv preprint doi: https://doi.org/10.1101/030718; this version posted November 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

overestimation/underestimation for all parameters is no more than by 100%, in all the models the approach was applied, except in the case of the amplitude parameter in the Ising model. In this case too, while the a parameter is dependent on both the true amplitude a_I as well as true size parameter ξ, the effect is to the extend of $n = .38 - 1.44$ for ξ =5–1000nm, the case relevant in the case of membrane protein clusters.

For the models in Table 1, this means that (1) the scaling is either independent of other parameters or only mildly dependent (2) the theoretical scaling due to the exponential approximation is within 100%, in contrast with the radius of maximal aggregation, which can be several times higher (technically upto ∞) depending on models and parameter values.

We validated this theoretical approach by means of Monte Carlo simulations. We simulated Gaussian cluster processes in a unit square for different conditions, such as varying the numbers of points per cluster as well as cluster radius. The empirical PCF of these point patterns were fitted to both the theoretical PCF for Gaussian point patterns, as well as the functional approximation $g_a(r)$, and the various parameters estimated. The estimates for N, the number of points per cluster is shown in Figure 5b. It can be seen that the simulations agree with the theoretical prediction, with estimates using $g_a(r)$ being overestimated, whereas the fit to Gaussian PCF providing accurate results.

Discussion

We have theoretically analyzed three spatial statistics based *model-free* methods for cluster parameters that have been proposed in the membrane protein imaging literature. They are: the radius of maximal aggregation based on Besag $L(r)-r$ function and the radius of maximal aggregation based on normalized K-function, both primarily estimators for cluster size, and the estimation based on the functional approximation with an exponential function for the Pair Correlation Function, proposed in the PC-PALM method. We were able to derive the theoretical relation between the radius of maximal aggregation and the true cluster parameters, for a diverse set of models, along with a theoretical lower bound for it. Our results illustrate the fact that the ratio of the radius of maximal aggregation (in the $L(r) - r$ case) to the true cluster size depends on the true cluster size as well as the number of clusters per unit area (or corresponding parameters, such as amplitude) for all the models considered. This dependence points to the difficulties of parameter identifiability using this technique, and also has implications in the interpretation of empirical $L(r) - r$ curves. In the case of the Pair Correlation approach, we were able to derive the scaling laws between the parameters of the approximate model and the true model, based on the Least Square Error criteria. From both the identifiability point of view as well as the scale of bias (between the true process parameters and the estimators), it appears that the Pair Correlation approach performs better, at least for the models our approach was applied on. While only a limited set of models were analyzed here, they show the limits of the estimators, and extending the analysis to other models is straightforward.

Also, the analysis shows that it might be possible to obtain theoretical bounds for parameters given a set of candidate models, e.g. by taking the worst bounds among candidates, even though the specific candidate model for a system is not known or is difficult to be inferred. It also points to a possible approach to reducing the bias: by using non-parameteric models for the *model-free* PCF, although care must be taken against overfitting and also in interpreting the results. This work only deals with the accuracy limits of the estimators, their precision could also be important

bioRxiv preprint doi: https://doi.org/10.1101/030718; this version posted November 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

in practical applications, which must be analyzed separately. The results presented in this work are not limited to membrane protein clusters, and are applicable to any system with spatial clustering.

Appendix

Lemma 0.1. *Let $h : \Re^+ \mapsto \Re^+$ be a unimodal differentiable function with a unique maximum at $r_m > 0$ and a derivative satisfying $h'(r) > 0$ for $0 \leq r < r_m$, and $h'(r) < 0$ for $r > r_m$. Note: this is satisfied by all the models in Table 1.*

Further assume that there exists $r^* > 0$ that satisfies

$$H(r^*) - r^* h(r^*) = 0.$$

Then the radius of maximal aggregation $r_a \geq r^*$ where r_a is obtained as a solution to (10) for some $A > 0$.

Furthermore as $A \to \infty$, we have $r_a \to r^*$.

Proof. Define

$$w(r) = H(r) - rh(r).$$

Clearly $w(0) = 0$ and the derivative satisfies $w'(r) = -rh'(r)$.

From the properties of h' we have $w'(r) \leq 0$ for $0 \leq r < r_m$, with strict inequality for $0 < r < r_m$, and $w'(r) > 0$ for $r > r_m$. Hence

$$w(r) < 0 \text{ for } 0 < r \leq r_m. \tag{14}$$

Since $w(r^*) = 0$ it follows that $r^* > r_m$. Moreover since $w'(r)$ is strictly positive for $r \in (r_m, r^*]$, it follows that $w(r) < 0$ for $r \in (r_m, r^*)$. Combining with (14) it follows that $w(r) < 0$ for $r \in (0, r^*)$.

Now, we know that r_a satisfies (10) for some $A > 0$. Thus we must have $w(r_a) > 0$ and hence it follows that $r_a \geq r^*$.

Now consider the situation in which $A \to \infty$. Define

$$z(r) = \frac{h(r)^2}{H(r) - rh(r)}$$

to denote the expression on the right hand side of (10) without the factor of 4π included. Since $z(r) = \frac{h(r)^2}{w(r)}$ we know from the earlier analysis of w that $z(r) \leq 0$ for $r < r^*$ and $z(r) \geq 0$ for $r < r^*$. Now consider the derivative of z. We have

$$\begin{aligned} z'(r) &= \frac{(H(r) - rh(r))2h(r)h'(r) + rh'(r)h(r)^2}{(H(r) - rh(r))^2} \\ &= \frac{2h(r)h'(r)H(r) - rh'(r)h(r)^2}{(H(r) - rh(r))^2} \\ &= \frac{h(r)h'(r)(2H(r) - rh(r))}{(H(r) - rh(r))^2} \end{aligned} \tag{15}$$

Now consider the function $q(r) = 2H(r) - rh(r)$ for $r \geq r^*$. At $r = r^*$ we have $q(r^*) = 2H(r^*) - r^*h(r^*) = H(r^*) > 0$. Moreover the derivative of this function is $q'(r) = h(r) - rh'(r)$ which is non-negative for $r > r^*$ because $h'(r) < 0$. Thus $q(r) > 0$ for $r > r^*$. This observation combined with the fact that $h'(r) < 0$ for $r > r^*$ and (15) implies that $z'(r) < 0$ for $r > r^*$. Thus we have that z is strictly decreasing in the interval (r^*, ∞). Moreover $z(r) \to \infty$ as r approaches r^* from above. Hence as $A \to \infty$ the left hand side of (10) $\to \infty$ and thus by virtue of (10) we must have $r_a \to r^*$. □

bioRxiv preprint doi: https://doi.org/10.1101/030718; this version posted November 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Author Contributions

A.S. conceived the project, performed the analysis, and wrote the manuscript, with inputs and guidance from J.U. and overall supervision from A.R. All authors read and approved the final manuscript.

Acknowledgments

A.S. thanks Prof. Ivo Sbalzarini, Hendrik Deschout and Dileep Kalathil for early discussions and guidance. This work was financially supported by Swiss National Science Foundation grants No. 200021–125319 and No. 20021–132206. A.S. was funded by a PhD fellowship grant from NCCBI. The authors declare that there are no conflicts of interest.

References

1. Sieber, J. J., K. I. Willig, C. Kutzner, C. Gerding-Reimers, B. Harke, G. Donnert, B. Rammner, C. Eggeling, S. W. Hell, H. Grubmuller, and T. Lang, 2007. Anatomy and dynamics of a supramolecular membrane protein cluster. Science 317:1072–1076.
2. Lang, T., and S. O. Rizzoli, 2010. Membrane protein clusters at nanoscale resolution: more than pretty pictures. Physiology 25:116–124.
3. Hartman, N. C., and J. T. Groves, 2011. Signaling clusters in the cell membrane. Curr Opin Cell Biol 23:370–376.
4. Garcia-Parajo, M. F., A. Cambi, J. A. Torreno-Pina, N. Thompson, and K. Jacobson, 2014. Nanoclustering as a dominant feature of plasma membrane organization. J Cell Sci 127:4995–5005.
5. Betzig, E., G. H. Patterson, R. Sougrat, O. W. Lindwasser, S. Olenych, J. S. Bonifacino, M. W. Davidson, J. Lippincott-Schwartz, and H. F. Hess, 2006. Imaging Intracellular Fluorescent Proteins at Nanometer Resolution. Science 313:1642–1645.
6. Hess, S. T., T. P. Girirajan, and M. D. Mason, 2006. Ultra-High Resolution Imaging by Fluorescence Photoactivation Localization Microscopy. Biophys J 91:4258 – 4272.
7. Rust, M. J., M. Bates, and X. Zhuang, 2006. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). Nat Methods 3:793–796.
8. Williamson, D. J., D. M. Owen, J. Rossy, A. Magenau, M. Wehrmann, J. J. Gooding, and K. Gaus, 2011. Pre-existing clusters of the adaptor Lat do not participate in early T cell signaling events. Nat Immunol 12:655–662.
9. Sengupta, P., T. Jovanovic-Talisman, D. Skoko, M. Renz, S. L. Veatch, and J. Lippincott-Schwartz, 2011. Probing protein heterogeneity in the plasma membrane using PALM and pair correlation analysis. Nat Methods 8:969–975.
10. Lillemeier, B. F., M. A. Mörtelmaier, M. B. Forstner, J. B. Huppa, J. T. Groves, and M. M. Davis, 2009. TCR and Lat are expressed on separate protein islands on T cell membranes and concatenate during activation. Nat Immunol 11:90–96.
11. Prior, I. A., C. Muncke, R. G. Parton, and J. F. Hancock, 2003. Direct visualization of Ras proteins in spatially distinct cell surface microdomains. J Cell Biol 160:165–170.
12. Scarselli, M., P. Annibale, C. Gerace, and A. Radenovic, 2013. Enlightening G-protein-coupled receptors on the plasma membrane using super-resolution photoactivated localization microscopy. Biochem Soc T 41:191–196.
13. Deschout, H., A. Shivanandan, P. Annibale, M. Scarselli, and A. Radenovic, 2014. Progress in quantitative Single Molecule Localization Microscopy. Histochem Cell Biol 142:5–17.
14. Rubin-Delanchy, P., G. L. Burn, J. Griffié, D. J. Williamson, N. A. Heard, A. P. Cope, and D. M.

bioRxiv preprint doi: https://doi.org/10.1101/030718; this version posted November 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Owen, 2015. Bayesian cluster identification in single-molecule localization microscopy data. Nat Methods .

15. Levet, F., E. Hosy, A. Kechkar, C. Butler, A. Beghin, D. Choquet, and J.-B. Sibarita, 2015. SR-Tesseler: a method to segment and quantify localization-based super-resolution microscopy data. Nat Methods .
16. Veatch, S. L., B. B. Machta, S. A. Shelby, E. N. Chiang, D. A. Holowka, and B. A. Baird, 2012. Correlation functions quantify super-resolution images and estimate apparent clustering due to over-counting. PLOS ONE 7:e31457.
17. Shivanandan, A., H. Deschout, M. Scarselli, and A. Radenovic, 2014. Challenges in quantitative Single Molecule Localization Microscopy. FEBS Lett 588:3595–3602.
18. Shivanandan, A., J. Unnikrishnan, and A. Radenovic, 2015. Accounting for Limited Detection Efficiency and Localization Precision in Cluster Analysis in Single Molecule Localization Microscopy. PLOS ONE 10:e0118767.
19. Kiskowski, M. A., J. F. Hancock, and A. K. Kenworthy, 2009. On the Use of Ripley's K-Function and Its Derivatives to Analyze Domain Size. Biophys J 97:1095–1103.
20. Antoku, Y., P. Dedecker, P. S. Pinheiro, T. Vosch, and J. B. Sørensen, 2015. Spatial distribution and temporal evolution of DRONPA-fused SNAP25 clusters in adrenal chromaffin cells. Photochem Photobiol Sci 14:1005–1012.
21. Roy, N. H., J. Chan, M. Lambelé, and M. Thali, 2013. Clustering and Mobility of HIV-1 Env at Viral Assembly Sites Predict Its Propensity To Induce Cell-Cell Fusion. J Virol 87:7516–7525.
22. Pezzarossa, A., 2012. Imaging plasma membrane domains in signal-transduction pathways. Ph.D. thesis, Leiden University.
23. Malkusch, S., W. Muranyi, B. Müller, H.-G. Kräusslich, and M. Heilemann, 2013. Single-molecule coordinate-based analysis of the morphology of HIV-1 assembly sites with near-molecular spatial resolution. Histochem Cell Biol 139:173–179.
24. Lagache, T., G. Lang, N. Sauvonnet, and J.-C. Olivo-Marin, 2013. Analysis of the Spatial Organization of Molecules with Robust Statistics. PLOS ONE 8:e80914.
25. Parton, R. G., and J. F. Hancock, 2004. Lipid rafts and plasma membrane microorganization: insights from Ras. Trends Cell Biol 14:141–147.
26. Felizzi, F., and D. Iber, 2014. Integrin clustering as a result of local membrane deformations and local signaling feedbacks. Physica A 408:198–211.
27. Gurry, T., O. Kahramanogullari, and R. G. Endres, 2009. Biophysical mechanism for ras-nanocluster formation and signaling in plasma membrane. PLOS ONE 4:e6148.
28. Tobin, S. J., E. E. Cacao, D. W. W. Hong, L. Terenius, V. Vukojevic, and T. Jovanovic-Talisman, 2014. Nanoscale effects of ethanol and naltrexone on protein organization in the plasma membrane studied by Photoactivated Localization Microscopy (PALM). PLOS ONE 9:e87225.
29. Cisse, I. I., I. Izeddin, S. Z. Causse, L. Boudarene, A. Senecal, L. Muresan, C. Dugast-Darzacq, B. Hajj, M. Dahan, and X. Darzacq, 2013. Real-Time Dynamics of RNA Polymerase II Clustering in Live Human Cells. Science 341:664–667.
30. Albertazzi, L., D. van der Zwaag, C. M. Leenders, R. Fitzner, R. W. van der Hofstad, and E. Meijer, 2014. Probing exchange pathways in one-dimensional aggregates with super-resolution microscopy. Science 344:491–495.
31. You, C., C. P. Richter, S. Lochte, S. Wilmes, and J. Piehler, 2014. Dynamic submicroscopic signaling zones revealed by pair correlation tracking and localization microscopy. Anal Chem 86:8593–8602.
32. Pertsinidis, A., K. Mukherjee, M. Sharma, Z. P. Pang, S. R. Park, Y. Zhang, A. T. Brunger, T. C. Südhof, and S. Chu, 2013. Ultrahigh-resolution imaging reveals formation of neuronal SNARE/Munc18 complexes in situ. Proc Natl Acad Sci U S A 110:E2812–E2820.
33. Veatch, S. L., P. Cicuta, P. Sengupta, A. Honerkamp-Smith, D. Holowka, and B. Baird, 2008. Critical

bioRxiv preprint doi: https://doi.org/10.1101/030718; this version posted November 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

fluctuations in plasma membrane vesicles. ACS Chem Biol 3:287–293.
34. Machta, B. B., S. Papanikolaou, J. P. Sethna, and S. L. Veatch, 2011. Minimal model of plasma membrane heterogeneity requires coupling cortical actin to criticality. Biophys J 100:1668–1677.
35. Shivanandan, A., A. Radenovic, and I. F. Sbalzarini, 2013. MosaicIA: an ImageJ/Fiji plugin for spatial pattern and interaction analysis. BMC Bioinformatics 14:349.
36. Sbalzarini, I. F., A. Mezzacasa, A. Helenius, and P. Koumoutsakos, 2005. Effects of organelle shape on fluorescence recovery after photobleaching. Biophys J 89:1482–1492.
37. Møller, J., and R. Waagepetersen, 2004. Statistical Inference and Simulation for Spatial Point Processes. Chapman and Hall/CRC, Boca Raton.
38. Diggle, P. J., 2003. Statistical Analysis of Spatial Point Patterns. Arnold, London.
39. Illian, J., A. Penttinen, H. Stoyan, and D. Stoyan, 2008. Statistical analysis and modelling of spatial point patterns, volume 70. John Wiley & Sons.
40. Lee, S.-H., J. Y. Shin, A. Lee, and C. Bustamante, 2012. Counting single photoactivatable fluorescent molecules by photoactivated localization microscopy (PALM). Proc Natl Acad Sci U S A .
41. Annibale, P., S. Vanni, M. Scarselli, U. Rothlisberger, and A. Radenovic, 2011. Quantitative Photo Activated Localization Microscopy: Unraveling the Effects of Photoblinking. PLOS ONE 6.
42. Greenfield, D., A. L. McEvoy, H. Shroff, G. E. Crooks, N. S. Wingreen, E. Betzig, and J. Liphardt, 2009. Self-Organization of the Escherichia coli Chemotaxis Network Imaged with Super-Resolution Light Microscopy. PLOS Biol 7:e1000137.
43. Baddeley, A. J., and R. Turner, 2004. Spatstat: An R Package for Analyzing Spatial Point Pattens.
44. Sengupta, P., T. Jovanovic-Talisman, and J. Lippincott-Schwartz, 2013. Quantifying spatial organization in point-localization superresolution images using pair correlation analysis. Nat Protoc 8:345–354.
45. Ghorbani, M., 2013. Cauchy cluster process. Metrika 76:697–706.
46. Jalilian, A., Y. Guan, and R. Waagepetersen, 2013. Decomposition of Variance for Spatial Cox Processes. Scand J Stat 40:119–137.

bioRxiv preprint doi: https://doi.org/10.1101/030718; this version posted November 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Model (r_t)	$g(r)-1$	$K(r)-\pi r^2$
Gaussian (σ) (39)	$\frac{1}{4\pi\kappa\sigma^2}\exp(\frac{-r^2}{4\sigma^2})$	$\frac{1}{\kappa}(1-\exp(\frac{-r^2}{4\sigma^2}))$
disk (R) (39)	$\frac{2}{\pi^2R^2\kappa}(\cos^{-1}(\frac{r}{2R})-\frac{r}{2R}\sqrt{1-\frac{r^2}{4R^2}})$	†
Cauchy (ω)(45)	$\frac{1}{8\pi\omega^2\kappa}(1+\frac{r^2}{4\omega^2})^{-3/2}$	$\frac{1}{\kappa}(1-\frac{1}{\sqrt{1+\frac{r^2}{4\omega^2}}})$
variance Gamma $\nu=1/2$ (η) (46)	$\frac{1}{2\pi\eta^2\kappa}\exp(-r/\eta)$	$\frac{1}{\kappa}\left(1-e^{-\frac{r}{\eta}}\left(1+\frac{r}{\eta}\right)\right)$
Ising (16)	$a_Ir^{-1/4}\exp(-r/\xi)$	$2\pi a_I\xi^{7/4}\left(\Gamma\left(\frac{7}{4}-\Gamma\left(\frac{7}{4},\frac{r}{\xi}\right)\right)\right)$

Table 1. Cluster models used for analysis. † $\frac{2}{\kappa\pi}(\frac{r^2\cos^{-1}\left(\frac{r}{2R}\right)}{R^2}-\frac{r\sqrt{1-\frac{r^2}{4R^2}}\left(r^2+2R^2\right)}{4R^3}+\sin^{-1}\left(\frac{r}{2R}\right))$. Also, for disk model, the functions provided here are for $r\leq 2R$, for $r>2R$, it is 0. Note that for disk, $g(r)=1$ at $r\geq 2R$, which provides a simple estimator for R.

Cluster model	Expression for $p=r_a/r_t$	Theoretical lower bound for p (to 5 digits)
Gaussian ($p=r_a/\sigma$)	$\kappa\sigma^2=\frac{e^{-\frac{p^2}{4}}p^2}{8\pi\left(-p^2+2e^{\frac{p^2}{4}}-2\right)}$	2.24181
Disk ($p=r_a/R$)	$\kappa R^2=\frac{p^2\left(p\sqrt{4-p^2}-4\arccos\left(\frac{p}{2}\right)\right)^2}{\pi^2\left(\sqrt{4-p^2}(3p^2-2)p-8p^2\arccos\left(\frac{p}{2}\right)+8\arcsin\left(\frac{p}{2}\right)\right)}$	1.29564
Cauchy($p=r_a/\omega$)	$\kappa\omega^2=\frac{p^2}{\pi(p^2+4)^{3/2}\left((p^2+4)^{3/2}-4p^2-8\right)}$	2.54404
varGamma ($p=r_a/\eta$)	$\kappa\eta^2=\frac{p^2}{4\pi(\exp(2p)-\exp(p)(p^2+p+1))}$	1.79328
Ising ($p=r_a/\xi$)	$\frac{1}{2\pi}a_I^{-1}\xi^{1/4}=\frac{\exp(-2p)p^{3/2}}{4\pi\left(-\exp(-p)p^{7/4}-\Gamma\left(\frac{7}{4},p\right)+\Gamma\left(\frac{7}{4}\right)\right)}$	1.37220

Table 2. Exact expressions for the radius of maximal aggregation r_a for different cluster models.

Cluster model	$m=d/r_t$	$n=a/a_t$	$l=N_a/N_t$
Gaussian	1.54	1.26	1.48
disk	.8	1.81	1.48
Cauchy	1.7	1.17	.85
varGamma	1	1	1
Ising	.5	$2.15r_t^{-1/4}(\approx .38-1.44)$	.59

Table 3. Theoretical scaling for different cluster models, in using the exponential approximation for PCF and using Least Square Error criteria. d,a,N_a correspond to the approximate PCF model $g_a(r)=1+a\exp(-r/d)$. True parameters r_t,a_t and N_t corresponding to the model PCFs of the form $f(r)=a_tv(r,r_t)$ can be obtained from Table 1 and using 4. The minimum r_m value, used in the calculation of the Squared Error E (in 13), for each model is as follows: Gaussian - 6σ, disk - $3R$ and Ising - 4ξ, and higher values for r_m give the same results. In the case of Cauchy model $r_m=\infty$ was used, and for varGamma any $r_m>0$ corresponds to the results in the table. The $m_{.95}$ values: $m_{.95}=d/r_{.95}$, $r_{.95}$ being the scale at which 95% of points are expected to lie, are .63,.82,.38 and .28 respectively, for Gaussian, disk, Cauchy and varGamma models.

bioRxiv preprint doi: https://doi.org/10.1101/030718; this version posted November 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

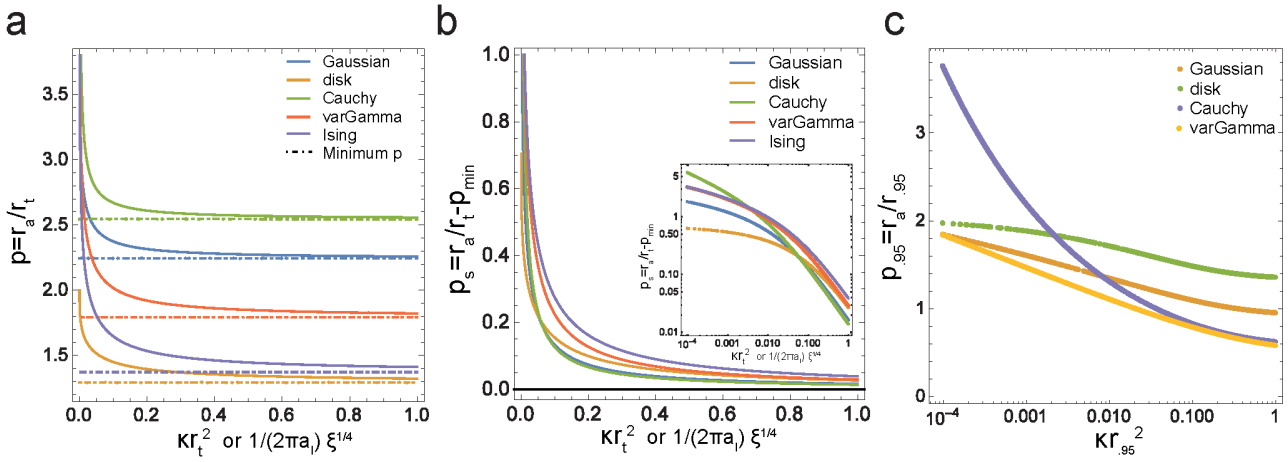

Figure 1. Relation between the radius of maximal aggregation and true cluster size.
a For different cluster models, the relation between the ratio of radius of maximal aggregation r_a and cluster size parameter of the true process r_t, as a function of the number of clusters per unit area κ and r_t. The minimum p value is obtained by exploiting the singularity in (10), given in Table 2 **b** Plots in **a** after translating by the minimum p and in log-log scale (inset). Note the partial power law like shape. **c** $p_{.95}$, the ratio between r_a and $r_{.95}$, the latter being the true scale within which 95% of all clustered points lie, plotted against $\kappa r_{.95}^2$. It can be seen that the relationships are model dependent. Note that for a sample with 10 clusters per μm^2 and $r_{.95} = 20nm$, $\kappa r_{.95}^2 = .004$.

bioRxiv preprint doi: https://doi.org/10.1101/030718; this version posted November 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

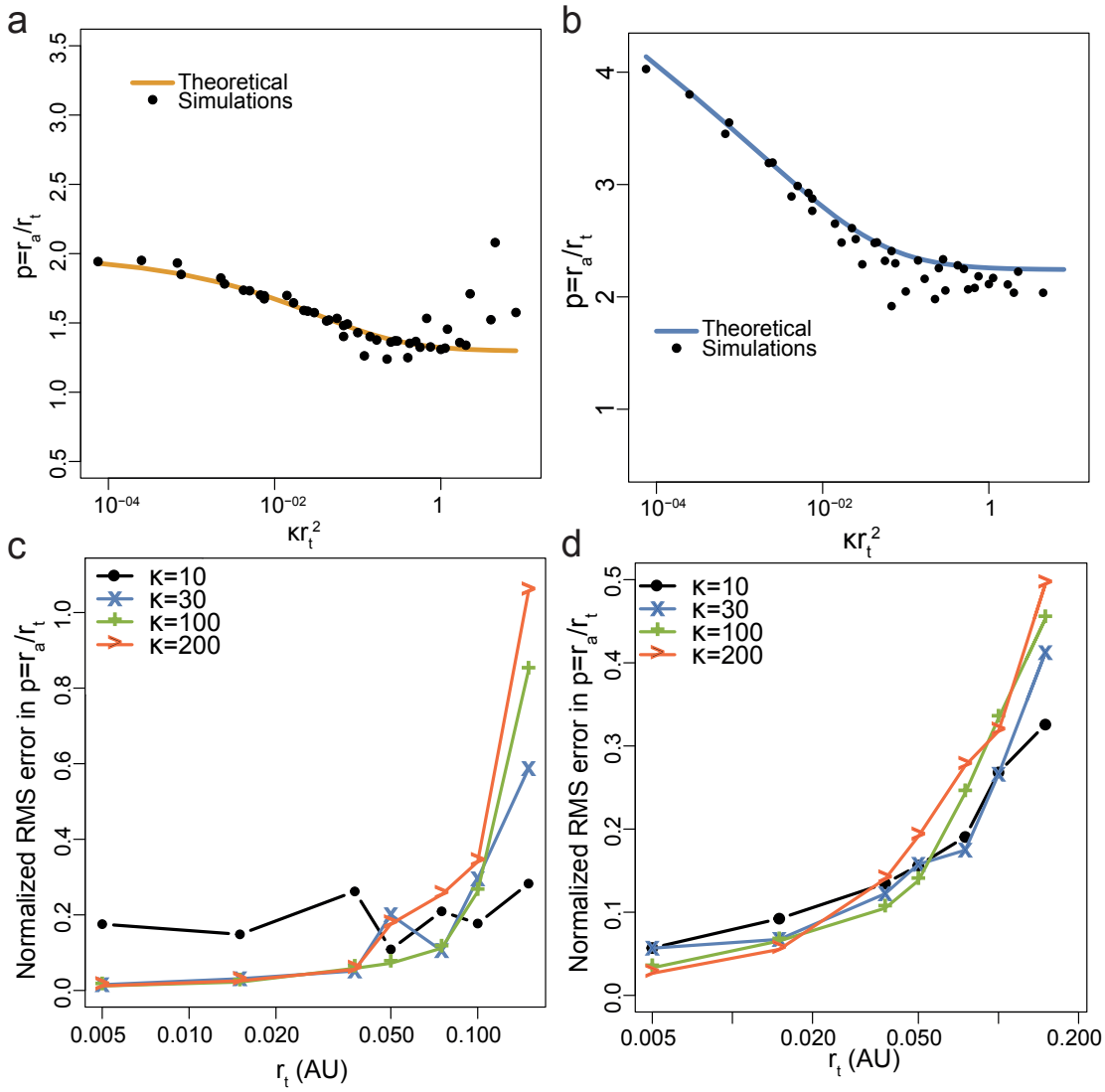

Figure 2. Comparison of theoretical results on $p = r_a/r_t$ with that from simulations **a,b** Results from theory (solid curve) as well as simulations on unit square window(dots), for disk and Gaussian clusters respectively. Only the mean value from 100 simulations are shown, for clarity, and the plot with error bars can be seen in Supporting Material Figure 1. It can be seen that in both disk and Gaussian cases, the mean values from simulations deviate from the theoretical values with increasing κr_t^2. **c,d** The Root Mean Squared error, normalized by the theoretical value, for disk and Gaussian clusters respectively, plotted against r_t. The colors denote different κ values. It can be seen that the error values are highly influenced by r_t.

bioRxiv preprint doi: https://doi.org/10.1101/030718; this version posted November 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

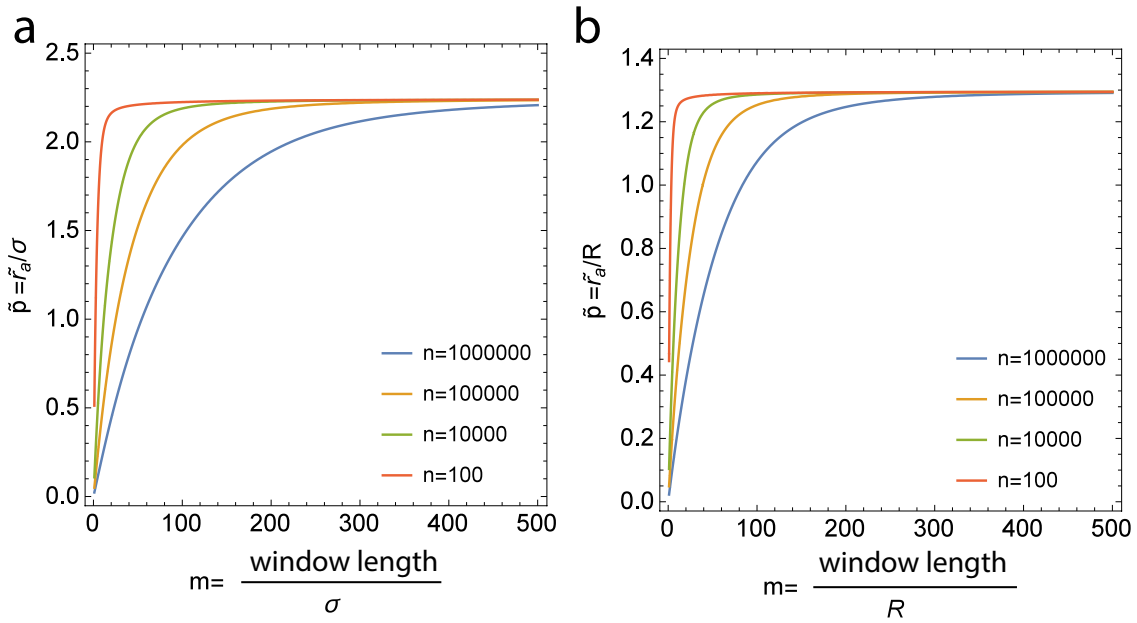

Figure 3. Results in the case of normalized K-function $\tilde{K}(r, n)$
a Gaussian clusters, **b** disk clusters. In the case of $\tilde{K}$, the $\tilde{p} = \tilde{r_a}/r_t$ depends on the number of points n and the ratio m between the size of the observation window (side length of a square in this case) and the true size parameter r_t, and converges to a maximum value at large m, which is approximately equal to the minimum values obtained in the case of r_a based on $L(r) - r$. Note that in the case of clusters with $R = 20nm$ with an area of analysis of size $10\mu m$, $m = 500$, and $\tilde{p}$ is close to the maximum value, and hence a constant. On the other hand, if the area of analysis was smaller, say $1\mu m$, $m = 50$, and $\tilde{p}$ depends on n.

bioRxiv preprint doi: https://doi.org/10.1101/030718; this version posted November 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

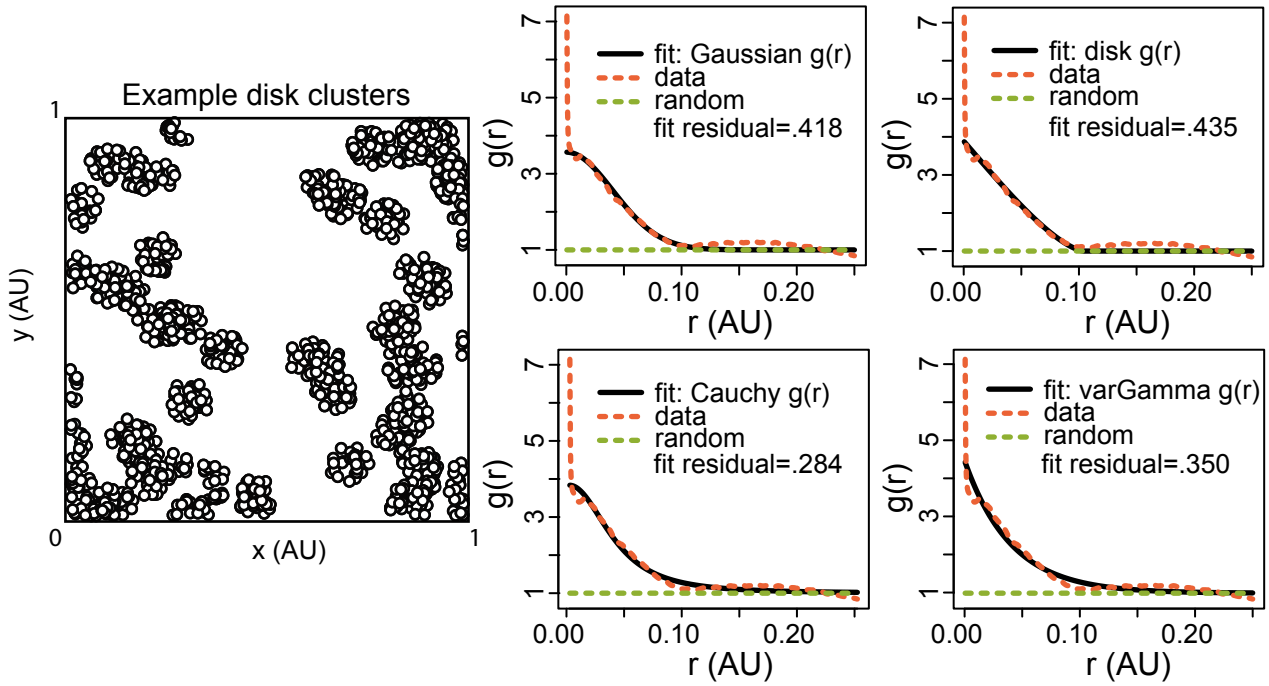

Figure 4. Demonstrative example of fitting model PCFs to the empirical PCF of a disk point pattern

The empirical PCF of the point pattern in the left is calculated, and is fit to the theoretical PCFs of various cluster processes. Fit results $(\hat{\kappa}, \hat{r_t}, \hat{\Delta})$, $\hat{\Delta}$ being the value of the objective function for the best fit parameters, called fit residual: Gaussian $(38.11, .028, .418)$, disk $(40.64, .052, .435)$, Cauchy $(21.55, .051, .284)$, varGamma $(27.86, .040, .350)$), whereas the true values of the disk point pattern are $(\kappa = 50, r_t = R = .05)$. Note that $\hat{r_t}$ is defined differently for different processes (Table 1). The Cauchy distribution is found to have the best fitness, whereas the disk one — the true model — has the worst. The $p = r_a/r_t$ corresponding to disk distribution, with the estimated parameters above is $\hat{p} = 1.44$. The maxima of $L(r) - r$ is at $\hat{r}_a = 0.072$, providing a $\hat{r}_t = \hat{r}_a/p = .05$, equal to the true R.

bioRxiv preprint doi: https://doi.org/10.1101/030718; this version posted November 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

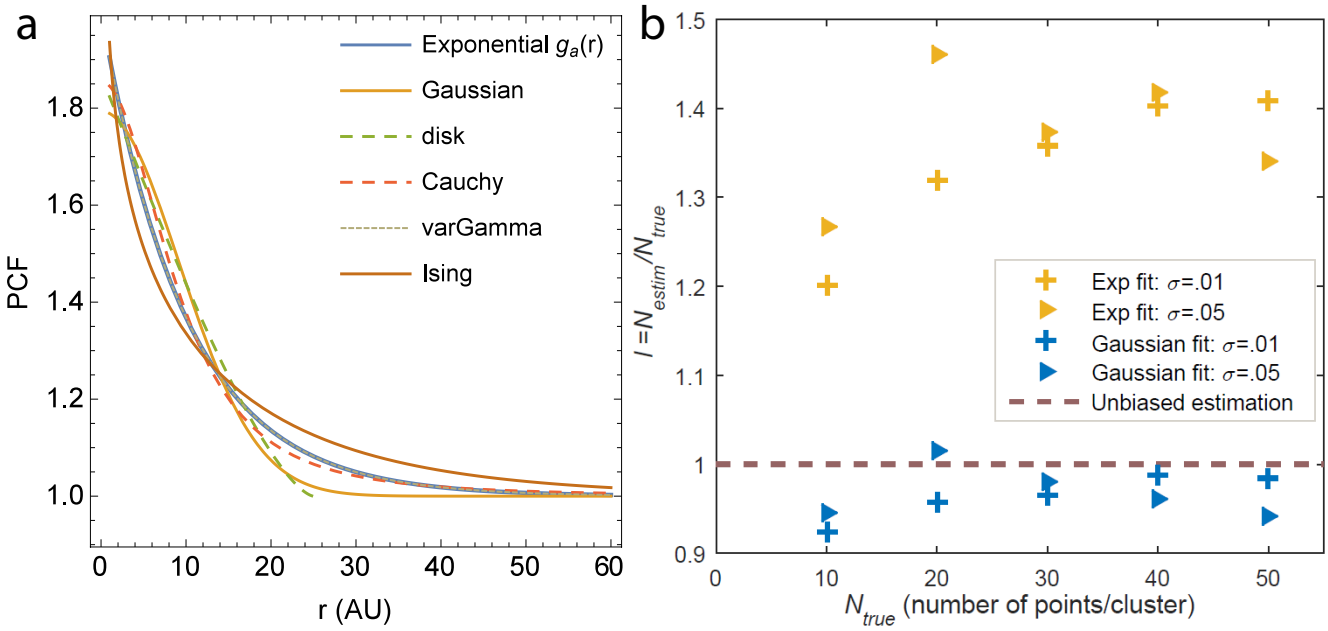

Figure 5. a Optimal Least Square Error fits for different models. For parameter values $a = 1$ and $d = 10$, the PCFs corresponding to different models in Table 1 is plotted, with the parameters scaled as per Table 3. For simplicity, only $r \geq 1$ is shown. **b** Mean estimates of N (number of points per cluster) from fitting the empirical PCF of Gaussian clustered point patterns with (1) Gaussian PCF (2) the exponential approximation $g_a(r)$(results from 20 simulations on a unit square window). The results broadly agree with the theoretical prediction of $l = 1.48$, approaching it with larger N_{true}. A plot with error bars can be seen in Supporting Material Figure 5.

Supporting Material for *On characterizing membrane protein clusters with model-free spatial correlation approaches*

A. Shivanandan, J. Unnikrishnan, A. Radenovic

Contents

1 Derivation of $f(p) = \kappa r_t^2$ and similar expressions for p **1**

2 95% scale for various models **1**

3 Radius of maximal aggregation in the case of $\tilde{K}(r, n)$ of Lagache et al **2**

4 Derivation for bias in PCF based on Least Squared Error **3**

5 Case of power law PCF **5**

1 Derivation of $f(p) = \kappa r_t^2$ and similar expressions for p

Here we derive the relation in the case of Neyman-Scott process with Gaussian shaped clusters. The derivation in the case of other distributions are similar, starting from the expressions in Table 1, Main Text.

We start from the K-function for Gaussian shaped clusters:

$$K(r) = \pi r^2 + \frac{1}{\kappa}(1 - \exp(\frac{-r^2}{4\sigma^2})). \tag{1}$$

In the form $K(r) = \pi r^2 + \frac{1}{A}H(r)$ as in Main Text, this corresponds to $A = \kappa$,$H(r) = 1 - \exp(\frac{-r^2}{4\sigma^2})$ and $h(r) = \frac{r}{2\sigma^2}\exp(\frac{-r^2}{4\sigma^2})$. Substituting in the equation

$$A = \frac{h(r_a)^2}{4\pi(H(r_a) - r_a h(r_a))} \tag{2}$$

from Main Text and rearrangement will give the relation as in Table 2, Main Text.

2 95% scale for various models

These were found by solving the CDF $\int_0^r f_{pdf}(r)dr = .95$ for r, where $f_{pdf}(r)$ is the radial probability density function for each model(1–3). In the case of Cauchy and varGamma models, marginal PDFs of r in polar coordinates were obtained from the bivariate PDFs in cartesian coordinates by standard transformation(multiplication by $2\pi r$). The results are given in the following table, along with the 95% limits. $K_\nu(.)$ denotes the modified Bessel function of the second kind.

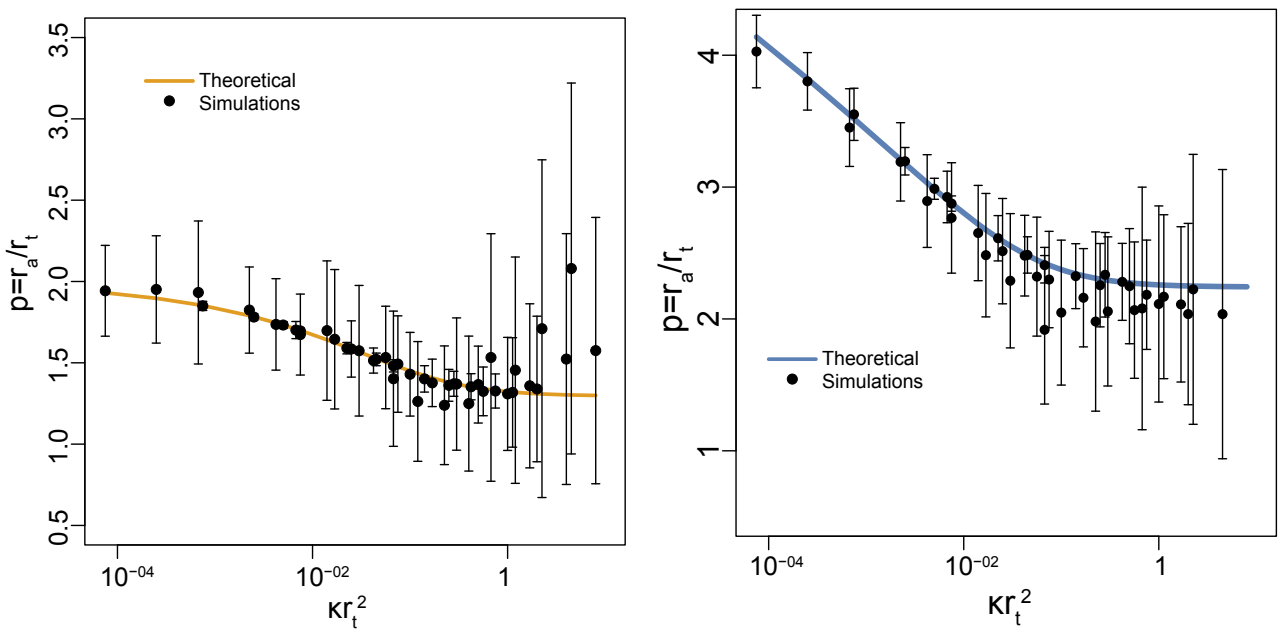

Figure 1: Comparison of $p = r_a/r_t$ from theory and simulations. Figure 1 in Main Text with error bars(σ).

Model	$f_{pdf}(r)$	$r_{.95} = u_{.95} r_t$	Lower bound for $p_{.95}$
Gaussian	$\frac{r}{\sigma^2}\exp\left(-\frac{r^2}{2\sigma^2}\right)$	2.448σ	.914
disk	$\frac{2r}{R^2}$	$.975R$	1.329
Cauchy	$\frac{r}{\omega^2}\left(1+\frac{r^2}{\omega^2}\right)^{-3/2}$	4.469ω	.568
VarGamma	$\frac{\sqrt[4]{2}r^{3/4}K_{-\frac{1}{4}}\left(\frac{r}{\eta}\right)}{\eta^{7/4}\Gamma\left(\frac{3}{4}\right)}$	3.547η	.505

3 Radius of maximal aggregation in the case of $\tilde{K}(r,n)$ of Lagache et al

Setting $\frac{\partial \tilde{K}(r,n)}{\partial r} = 0$ for disk clusters as discussed in Main Text, followed by routine manipulations lead us to the relation:

$$
\begin{aligned}
&-\frac{0.0210642p^2\left(\left(16-4p^2\right)\cos^{-1}(0.5p)+p\sqrt{4-p^2}\left(p^2-4\right)\right)\left(6.0286m^3+7.35489m^2p-18.9394mp^2+np^3\right)}{p^2-4}\\
&+0.00789906p\left(2.45163m^2-12.6263mp+np^2\right)\left(\sqrt{4-p^2}\left(p^2+2\right)p-8p^2\cos^{-1}\left(\frac{p}{2}\right)-8\sin^{-1}\left(\frac{p}{2}\right)\right)\\
&+0.0317468\left(m^3+1.22m^2p-3.14159mp^2+0.165876np^3\right)\left(\sqrt{4-p^2}\left(p^2+2\right)p-8p^2\cos^{-1}\left(\frac{p}{2}\right)-8\sin^{-1}\left(\frac{p}{2}\right)\right)\\
&=0, \quad (3)
\end{aligned}
$$

where $p = \tilde{r}_a/R$, $m = side/R$ where $A = side^2, P = 4.side$.

The contour plot of p vs m, based on this expression, is shown in the Main Text, for different values of n.

In the case of Gaussian clusters, the relation is simpler:

$$m^3\left(p^2-2e^{\frac{p^2}{4}}+2\right)+m^2p\left(1.22p^2-3.66e^{\frac{p^2}{4}}+3.66\right)+mp^2\left(-3.14159p^2+12.5664e^{\frac{p^2}{4}}-12.5664\right)$$
$$+np^3\left(0.165876p^2-0.82938e^{\frac{p^2}{4}}+0.82938\right)=0, \quad (4)$$

and the corresponding contour plot is provided in Main Text.

4 Derivation for bias in PCF based on Least Squared Error

We simply show the case for Ising model. Derivation for other models follow the same procedure. For $g_a(r) = 1 + a\exp(-r/d)$ and $f(r) = 1 + Ar^{-1/4}\exp(-r/D)$, the Least Squared Error criteria gives:

$$(\hat{a}, \hat{d}) = \arg\min_{a,d} E = \arg\min_{a,d} \int_0^{r_m} (f(r) - g_a(r))^2 dr. \quad (5)$$

We obtain: $E = -\frac{1}{2}a^2 d\left(-1+e^{-\frac{2r_m}{d}}\right) + \frac{A^2\sqrt{\frac{\pi}{2}}\sqrt{r_m}\mathrm{Erf}\left[\sqrt{2}\sqrt{\frac{r_m}{D}}\right]}{\sqrt{\frac{r_m}{D}}} - \frac{2aAr_m^{3/4}\left(\Gamma\left[\frac{3}{4}\right]-\Gamma\left[\frac{3}{4},\frac{(d+D)r_m}{dD}\right]\right)}{\left(\frac{(d+D)r_m}{dD}\right)^{3/4}}$

$$\frac{\partial E}{\partial a}=0 \implies \frac{\partial E}{\partial a} = -ad\left(-1+e^{-\frac{2r_m}{d}}\right) - \frac{2Ar_m^{3/4}\left(\Gamma\left[\frac{3}{4}\right]-\Gamma\left[\frac{3}{4},\frac{(d+D)r_m}{dD}\right]\right)}{\left(\frac{(d+D)r_m}{dD}\right)^{3/4}}=0$$

$$\frac{\partial E}{\partial d}=0 \implies \frac{\partial E}{\partial d} = -\frac{1}{2}a^2\left(-1+e^{-\frac{2r_m}{d}}\right) - \frac{a^2e^{-\frac{2r_m}{d}}r_m}{d} - \frac{2aAdDe^{-\frac{(d+D)r_m}{dD}}\left(\frac{r_m}{dD}-\frac{(d+D)r_m}{d^2D}\right)}{(d+D)r_m^{1/4}}$$
$$+\frac{3aAr_m^{3/4}\left(\frac{r_m}{dD}-\frac{(d+D)r_m}{d^2D}\right)\left(\Gamma\left[\frac{3}{4}\right]-\Gamma\left[\frac{3}{4},\frac{(d+D)r_m}{dD}\right]\right)}{2\left(\frac{(d+D)r_m}{dD}\right)^{7/4}}=0$$

Solving both equations separately for $a = \hat{a}$, we obtain:

$$\hat{a} = \frac{2Ae^{\frac{2r_m}{d}}r_m^{3/4}\left(\Gamma\left[\frac{3}{4}\right]-\Gamma\left[\frac{3}{4},\frac{(d+D)r_m}{dD}\right]\right)}{d\left(-1+e^{\frac{2r_m}{d}}\right)\left(\frac{(d+D)r_m}{dD}\right)^{3/4}}$$

and,

$$\hat{a} = \frac{\frac{4ADe^{-\frac{(d+D)r_m}{dD}}r_m^{3/4}}{d(d+D)} - \frac{3Ar_m^{7/4}\Gamma\left[\frac{3}{4}\right]}{d^2\left(\frac{(d+D)r_m}{dD}\right)^{7/4}} + \frac{3Ar_m^{7/4}\Gamma\left[\frac{3}{4},\frac{(d+D)r_m}{dD}\right]}{d^2\left(\frac{(d+D)r_m}{dD}\right)^{7/4}}}{-1+e^{-\frac{2r_m}{d}}+\frac{2e^{-\frac{2r_m}{d}}r_m}{d}}$$

Equating both the above expressions of $\hat{a}$, simplifying, and setting $m = d/D$ and $k = r_m/D$, we get:

$$\frac{2e^{\frac{2k}{m}}\left(\Gamma\left(\frac{3}{4}\right)-\Gamma\left(\frac{3}{4},k\left(1+\frac{1}{m}\right)\right)\right)}{e^{\frac{2k}{m}}-1} + \frac{me^{k\left(\frac{1}{m}-1\right)}\left(4\left(\frac{k}{m}+k\right)^{3/4}-3\Gamma\left(\frac{3}{4}\right)e^{\frac{k}{m}+k}+3e^{\frac{k}{m}+k}\Gamma\left(\frac{3}{4},k\left(1+\frac{1}{m}\right)\right)\right)}{(m+1)\left(m\left(e^{\frac{2k}{m}}-1\right)-2k\right)} = 0$$

Note that this equation does not contain the amplitude parameters a and A. A contour plot of this equation is shown in Figure 2. For reasonably large values of r_m (i.e., $r_m > 2D$), $m = \hat{d}/D = .5$. That is, the correlation length parameter estimated by the approximate model is half of the correlation length of the true model.

From these results, the parameter values $k = 4, m = .5$ (or any $k > 2$) can be substituted in the expression for $\hat{a}$, to obtain:

$$n = \frac{a}{A} = 2.15031D^{-1/4}$$

That is, the amplitude parameter of the approximate model is dependent on both the true amplitude parameter as well as the correlation length. The relationship is shown in Figure 3. This parameter could be $n = .38 - 1.44$ scaled from the true amplitude parameter for $D = 5 - 1000nm$, relevant scales for membrane protein clusters.

Now, the average number of points per cluster:

$$N_I = 1 + \rho \int_0^\infty (f(r) - 1) 2\pi r dr \approx 2\pi A D^{1.75} \Gamma\left(\frac{7}{4}\right)$$

$$N_a \approx 2\pi a d^2 \rho = 3.3777 A D^{1.75} = 0.584919 N_I$$

That is, the approximate model underestimates the average number of points per cluster by over 40%.

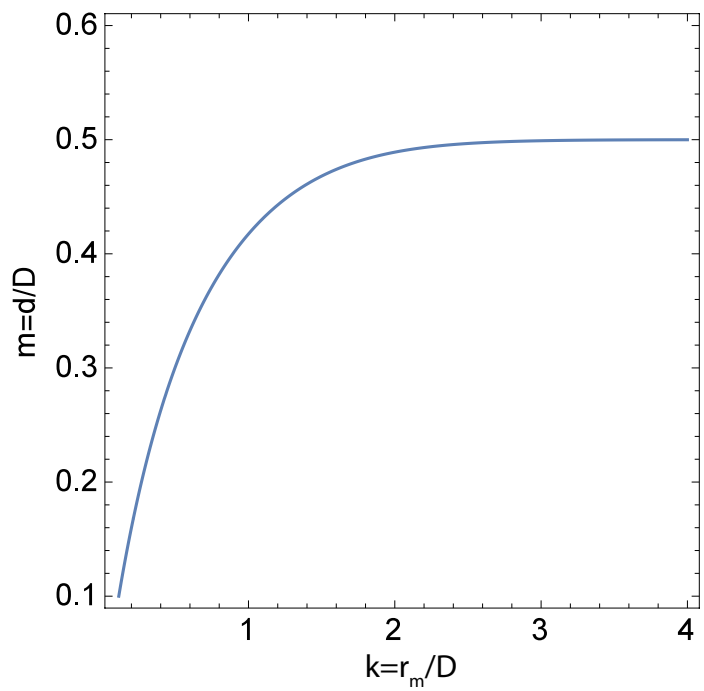

Figure 2: Contour plot of $k = r_m/D$ vs $m = d/D$ for Ising model. r_m is the distance value to which the Least Squares sum is taken. After $\approx r_m > 2D$, the m value is fixed at .5.

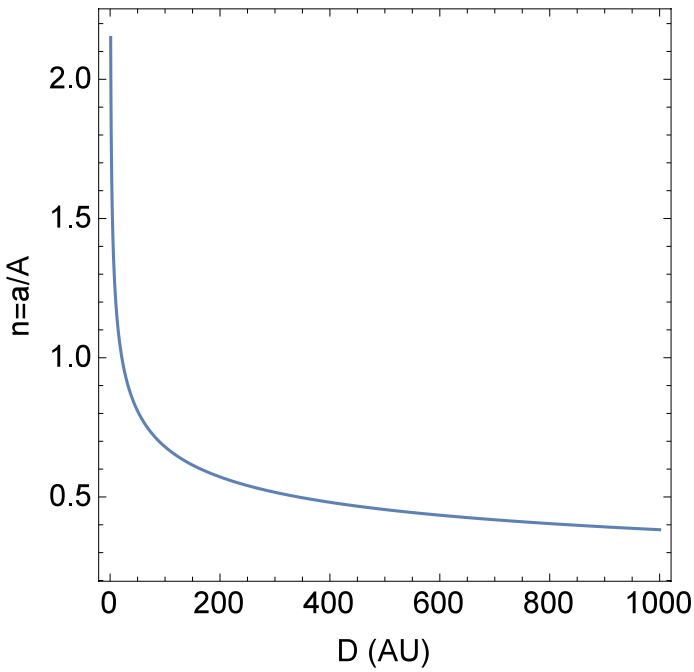

Figure 3: Plot of D vs $n = a/A$, at $k = 4, m = .5$.

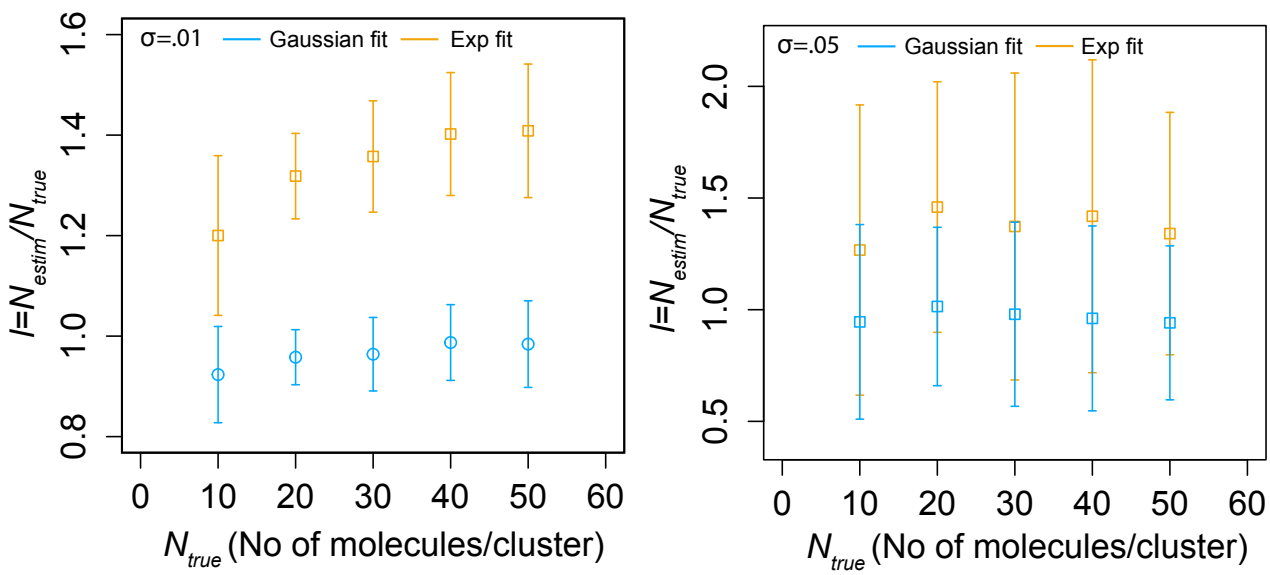

Figure 4: Comparison of fitting empirical PCF of Gaussian clusters to (1) exponential PCF g_a and (2) theoretical PCF of Gaussian clusters, for different true cluster σ. Figure 5b in Main Text shown with error bars(σ).

5 Case of power law PCF

In the case of the PCF $g(r) = 1 + c\left(\frac{r_0}{r}\right)^s$, assuming $s \neq 1$,

$$K(r) = \pi r^2 + \frac{2\pi c}{2-s}\left(\frac{r_0}{r}\right)^s r^2 \tag{6}$$

for $s < 2$.
A in (10) of Main Text will be $A = \frac{2-s}{2\pi c}$. Using (10), we get:

$$p = \frac{r_a}{r_0} = \left(\frac{c(2-s)}{2(s-1)}\right)^{1/s}. \tag{7}$$

A plot of this equation for different s is shown in Figure 5. It can be seen that p varies across orders of magnitude based on values of s and c.

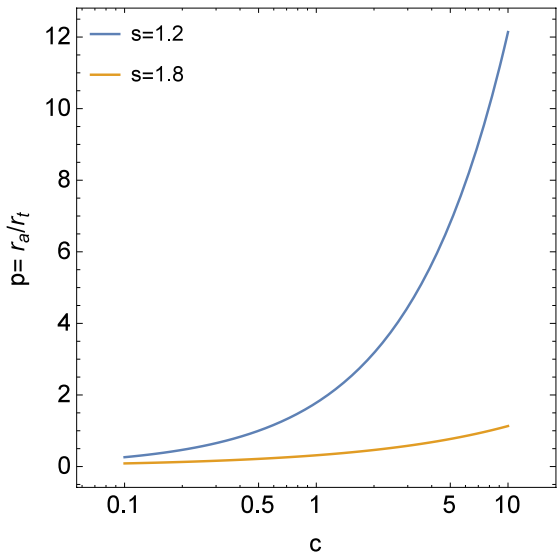

Figure 5: Ratio of radius of maximal aggregation to true cluster size parameter $p = \frac{r_a}{r_0}$ for power law PCF.

References

[1] Illian, J., A. Penttinen, H. Stoyan, and D. Stoyan, 2008. Statistical analysis and modelling of spatial point patterns, volume 70. John Wiley & Sons.

[2] Ghorbani, M., 2013. Cauchy cluster process. *Metrika* 76:697–706.

[3] Jalilian, A., Y. Guan, and R. Waagepetersen, 2013. Decomposition of Variance for Spatial Cox Processes. *Scand J Stat* 40:119–137.